Censored data is tricky. Suppose you are interested in studying survival following heart transplant surgery. A comparison group might be similarly sick patients who do not undergo transplant surgery. All other things being equal, do surgically treated patients have longer survival times than non-surgically treated patients? How to proceed? One approach might be to do a logistic regression analysis with outcome defined as 0/1 occurrence of death by 1 year. Or 5 years. Or 10 years. A limitation of this approach is the possibility of loss to follow-up. At the end of your study, some study participants will have died. Others will have been lost to follow up. And still others will be known to be alive at last contact. In some instances you have complete information (eg; a study subject is known to have died at 2.3 years post transplant). In other instances, only partial information is known (eg; another study subject is known to have survived 1.8 years but there is no additional information). Data such as these are known as survival data and special techniques are required for their analysis. Fortunately, they exist! They have the advantage of taking into consideration the available information on every subject (so much better than tossing these observations out!).

The survival analyses introduced in this unit are used to address questions such as the following:

1. What is the estimated probability of subject surviving a specified amount of time? (eg; what is the five-year survival rate?)

2. What is the comparative survival experience of two independent groups of subjects? (eg; relative to standard care, is heart transplant surgery associated with a statistically significant improvement in survival?)

3. Among possibly multiple indicators of risk (eg – age, comorbidities), which are statistically significantly associated with greater hazard of event (eg; what are the risk factors for poor prognosis following heart transplant surgery?)
Table of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Learning Objectives ..................................................</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction and Examples ..........................................</td>
<td>4</td>
</tr>
<tr>
<td>2.</td>
<td>Example 1 – Survival Following a Heart Attack ..................</td>
<td>11</td>
</tr>
<tr>
<td>3.</td>
<td>Notation and Definitions ............................................</td>
<td>14</td>
</tr>
<tr>
<td>4.</td>
<td>Probability Models for Survival Data ................................</td>
<td>19</td>
</tr>
<tr>
<td>5.</td>
<td>The Kaplan-Meier Curve - Model Free Estimation ..................</td>
<td>22</td>
</tr>
<tr>
<td>6.</td>
<td>The Log Rank and Related Tests - Model Free Comparison .......</td>
<td>30</td>
</tr>
<tr>
<td>7.</td>
<td>Introduction to the Cox PH Model ...................................</td>
<td>37</td>
</tr>
<tr>
<td>8.</td>
<td>Interpretation of a Cox PH Model ...................................</td>
<td>39</td>
</tr>
<tr>
<td>9.</td>
<td>Hypothesis Testing Using the Cox PH Model .......................</td>
<td>43</td>
</tr>
<tr>
<td>10.</td>
<td>Evaluating the Proportional Hazards Assumption ..................</td>
<td>45</td>
</tr>
<tr>
<td>11.</td>
<td>Regression Diagnostics for the Cox PH Model .....................</td>
<td>48</td>
</tr>
</tbody>
</table>

Appendix
Overview of Maximum Likelihood Estimation of a Cox PH Model .............. 50
Learning Objectives

When you have finished this unit, you should be able to:

- Explain “time-to-event” data and provide examples.
- Define censoring and explain the three kinds of censoring: right censored, left censored and interval censored.
- Calculate Kaplan-Meier estimates of survival probabilities for a single sample of time-to-event data with right censoring.
- Draw a Kaplan-Meier curve of estimated survival probabilities for a single sample of time-to-event data with right censoring.
- State the null hypothesis of the log-rank test.
- Perform and interpret the log-rank test for the comparison of the survival experience of two independent groups in the setting of right censoring.
- Explain the idea of the hazard ratio and its similarity to the idea of relative risk.
- Define the Cox Proportional Hazards (PH) model.
- Extract point and confidence interval estimates of relative hazard (hazard ratio) from a fitted Cox PH model.
- Interpret the results of a Cox PH model analysis that examines the nature and significance of possibly multiple predictors of survival.
1. Introduction and Examples

The type of data that is of interest here is different from those that we have considered previously.

- In BIOSTATS 540: A sample of observations of a continuous variable (e.g. blood pressure, cholesterol) that is distributed Normal.

- In BIOSTATS 540: Two independent samples of observations of a continuous variable (e.g. blood pressure, cholesterol), one from each of two groups (the groups might have been males and females or controls and experimentals) that are distributed Normal.

- In BIOSTATS 540: One or two observations of discrete (in particular, count) data that is distributed Binomial (e.g. # heads in several tosses of a coin, # remissions of cancer among several persons treated with a new cancer therapy).

- In BIOSTATS 640: Paired observations of two discrete traits (e.g. race/ethnicity and religious affiliation) each of which has multiple possibilities (e.g. race/ethnicity might be coded as African/American, Latino, Asian, Other and religious affiliation might be coded as Muslim, Hindu, Judao/Christian, Other) and which we analyze using contingency table approaches.

- In BIOSTATS 640: Observations of a normally distributed variable (e.g. blood pressure, cholesterol) which we investigate in relationship to a collection of hypothesized predictors (e.g. age, sex, health behaviors) using multivariable normal theory regression techniques (Unit 5).

- In BIOSTATS 640: Observations of a Bernoulli distributed binary discrete variable (e.g. “yes/no” disease) which we investigate in relationship to a collection of hypothesized predictors (e.g. exposure, age, sex, health behaviors) using multivariable logistic regression techniques (Unit 7).
Consider the following settings.

- A cancer study examines the time from onset of therapy to death. The goal is a descriptive one directed to an understanding of prognosis.

- A study of treatments for cardiovascular disease compares bypass surgery, angioplasty and medical therapy by examining the time from treatment until death. The setting is a randomized controlled trial with the objective of assessing the relative benefits of three alternative management approaches.

- A health services researcher might seek a description of the patterns of time from enrollment in a health plan to first utilization of services. The setting is health services planning.

In these settings, the focus is on a special type of continuous variable known as “time to event” data.

- “Time to event” data are such things as
  - length of time unemployed - measured from date of layoff
  - lifetime of a light bulb (“failure time” data).
  - elapsed time to death following diagnosis of disease (“survival time” data).

- A characteristic of “time to event” data is that it may (and often does) include observations that are incomplete. In particular, we may not know the actual time to event for every person.

  - For some individuals, we have actual event times.
  - For others, we know only the occasion at which observation ended and that the actual event time is some larger value (right censoring).
  - For still others, we might know only that the actual event time occurred previously (left censoring) or at some point within a window of time (interval censoring).

- Therefore, appropriate analysis must accommodate the mixture of complete (event time is known) and incomplete (event time is known only partially).

- Note: Other disciplines use the term “reliability theory” where we say “survival theory”.

Example #1 (Description) – What is the clinical course of lung cancer?

Interest is in understanding the clinical course of patients newly diagnosed with lung cancer:
Assemble cohort “with disease”

→

Follow forward in time

→

Report occasions of complications, death, etc.

This example illustrates two issues in a survival analysis:

1. The meaning of “disease”; and
2. The concept of “time zero”?

Example #2 (Hypothesis testing) – Evaluation of a new treatment for leukemia.

Is treatment with 6-mercaptopurine statistically significantly more effective in prolonging remission than standard care?

Assemble a cohort of “like” individuals

→

Randomize

Standard Care 6-Mercaptopurine

Follow over time

→

Follow over time

Compare standard care versus 6-mercaptopurine groups with respect to post-treatment duration of remission.
This example also illustrates the following issues of survival analyses:

3. **Summarization of “survival patterns”; and**
4. **Comparison of groups**

**Example #3 (multivariable modeling) – Exploring the nature of an economic recession.**

Economic recessions are hard. What factors are related to length of time unemployed? How might we conduct this study?

Assemble a cohort of “laid off” individuals
Compile (hypothesized) relevant information from welfare offices

Outcome: Length of Time Unemployed

Hypothesized Predictors:

- Proportion labor force out of work
- Previous wage
- Age at “laid off”
- Number of years employed at “laid off” job
- Etc.

Investigate separate and joint effects of the hypothesized correlates of “length of time unemployed”

In this example, the term “survival” is a **misnomer**, since it is referring to the length of time an individual is without a job. Nevertheless, the tools of survival analysis are appropriate for analyzing data of this sort. This example illustrates the issue of **multivariable model development in survival analysis**.
There are at least four (4) goals of a “time to event” analysis.

1. **Description** - To describe the pattern (distribution) of “time-to-event” times
   - Single population
   - “Life-table”, “Kaplan-Meier”
   - E.g. – Example #1: Lung cancer

2. **Hypothesis Testing** - To compare patterns of “time-to-event” across groups
   - Experiment, homogeneous groups
   - “Log Rank Test”
   - E.g. – Example #2: Experimental Treatment for leukemia

3. **Exploration** - To explore the influences of possibly several factors on “time to event”
   - Observational study or experiment
   - “Cox regression”, “Proportional Hazards” model
   - E.g. – Example #3: Recession

4. **To identify limitations in interpretation**
   - Biases
   - Precision

**Note:** The term “survival analysis” will be used in the pages that follow, instead of “time to event” analysis. “Survival analysis” will refer generally to time to event analysis, even when the outcome is different than death and may even be something desirable (e.g. new employment).
Survival Analysis Methodology addresses some unique issues, among them:

1. “At risk”. This needs to be defined for each survival analysis setting.

   An “at risk” group is a collection of “like” individuals who are similar with respect to everything (they have similar “profiles” on the explanatory variables) except the one (or more) factor(s) of interest in the analysis.

2. Time Zero

   Time zero is the start of follow-up and, as such, is the same for all individuals. Often, this reference time is NOT a calendar date. It is typically something like the occasion of diagnosis of disease, or the occasion of entry into study.

3. Incomplete data. As previously mentioned, sometimes (often, actually), we do not know the “time to event” values for every person. These observations of time to event are called censored data.

   Censored data, notwithstanding its incompleteness, is still useful because, for each individual, we at least know a portion of their survival time. In the case of right censoring, we know that his or her time to event is at least as long as the time we observed that person. Note- “Left censoring” and “interval censoring” are two other types of censoring; these are not addressed here.

4. Examining survival data at a single point in time is not meaningful.

   The picture on the following page illustrates why.
Illustration – Comparisons of Survival Rates Calculated at a Single Point in Time are Misleading.

In this illustration, patients are followed over time following the occurrence of a heart attack (time zero = occasion of the heart attack (MI)). Two groups are compared.

- Patients with normal blood pressure at the time of MI
- Patients with low blood pressure at the time of MI

- **Time “A”**: Low blood pressure at the time of MI is a risk factor for early death. The proportion still alive in the LOW BP group at time “A” is *less* than that in the NORMAL BP group.

- **Time “B”**: Later, however, low blood pressure confers a survival advantage. At time “B”, the proportion still alive is *greater* than that in the NORMAL BP group. A flip →

- **Thus, factors related to survival**: (1) may not be apparent at one point in time; and/or (2) may change over time.
2. Example 1
Survival Following a Heart Attack

Setting.

It has been suggested that low blood pressure is a predictor of shorter survival time following a heart attack, compared to normal blood pressure. For illustration purposes, consider a little case study of 13 patients hospitalized for heart attack (myocardial infarction). Each patient is followed over time to the occurrence of either death or conclusion of the study, whichever comes first. The goal of the analysis is to assess the role of low blood pressure at the time of the heart attack (MI) as a risk factor for early death.

There are two groups.
- Low BP Group: n=7
- Normal BP Group: n=6

Study Design.

- The study began on September 12, 1991.
- Follow-up ended on June 12, 2011.
- Patients were enrolled over time during this period.
- **Time zero** for each patient, is his or her date of heart attack.
- The value of \( X \) for each patient is their elapsed time between time zero and death or last follow-up, whichever occurred first.

Study Data

**LOW BP Group (n=7)**

<table>
<thead>
<tr>
<th>Patient</th>
<th>( X ) = Number Days Followed</th>
<th>Final Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>413</td>
<td>Dead</td>
</tr>
<tr>
<td>3</td>
<td>1075</td>
<td>Dead</td>
</tr>
<tr>
<td>5</td>
<td>1801</td>
<td>Dead</td>
</tr>
<tr>
<td>7</td>
<td>3044</td>
<td>Dead</td>
</tr>
<tr>
<td>8</td>
<td>3351</td>
<td>Dead</td>
</tr>
<tr>
<td>9</td>
<td>5551</td>
<td>Dead</td>
</tr>
<tr>
<td>10</td>
<td>6277</td>
<td>Dead</td>
</tr>
</tbody>
</table>

**NORMAL BP Group (n=6)**

<table>
<thead>
<tr>
<th>Patient</th>
<th>( X ) = Number Days Followed</th>
<th>Final Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>701</td>
<td>Alive</td>
</tr>
<tr>
<td>4</td>
<td>1735</td>
<td>Dead</td>
</tr>
<tr>
<td>6</td>
<td>2989</td>
<td>Dead</td>
</tr>
<tr>
<td>11</td>
<td>7293</td>
<td>Alive</td>
</tr>
<tr>
<td>12</td>
<td>7352</td>
<td>Alive</td>
</tr>
<tr>
<td>13</td>
<td>7434</td>
<td>Alive</td>
</tr>
</tbody>
</table>
Time to Event is Not the Same as Calendar Time

Illustration for Four (4) Subjects

Key:

X - Entered study
O - Alive at last observation
▲ - Dead at last observation

<table>
<thead>
<tr>
<th>Patient</th>
<th>Entered</th>
<th>Survived</th>
<th>Died/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9/12/91</td>
<td>413</td>
<td>day 413</td>
</tr>
<tr>
<td>2</td>
<td>7/11/09</td>
<td>701</td>
<td>day 701</td>
</tr>
<tr>
<td>3</td>
<td>2/5/92</td>
<td>1075</td>
<td>day 1075</td>
</tr>
<tr>
<td>13</td>
<td>1/16/92</td>
<td>7434</td>
<td>day 7434</td>
</tr>
</tbody>
</table>

Nature  Population/ Sample  Observation/ Data  Relationships/ Modeling  Analysis/ Synthesis
Time to Event is Not the Same as Calendar Time

Calendar Time versus Time to Event

The following is taken from an introduction to survival analysis written by Mark Lunt, at University of Manchester, England.

In both panels, the horizontal axis is scaled in units of months.

Calendar Time Panel
Horizontal axis depicts actual calendar time. →
Time zero = calendar date of enrollment.

Time on Study Panel
Horizontal axis depicts elapsed time. →
Time zero = start of follow up

Source: [http://personalpages.manchester.ac.uk/staff/mark.lunt/stats/10_Survival/handout.pdf](http://personalpages.manchester.ac.uk/staff/mark.lunt/stats/10_Survival/handout.pdf)
3. Notation and Definitions
(right censoring is assumed)

An individual’s survival data are described by a pair of variables, \((X, C)\). We are especially interested in the true survival time which is another random variable called \(T\).

- \(X\) – The observed time on study. Eg - for patient #1, \(X=413\), for patient #2, \(X=701\) and so on.

- \(C\) – An indicator variable, specifying whether or not the event of interest occurred while on study. Typically,
  - \(C=1\) indicates that the event DID occur Eg - for patient #1, \(C=1\)
  - \(C=0\) indicates that the event did NOT occur Eg - for patient #2, \(C=0\)

- \(T\) – The true time to event, sometimes observable, sometimes only partially observable. Here, because of right censoring, we have:
  - The event DID occur: Then \(C=1\) and \(T=X\) Eg - for patient #1, \(T=X=413\)
  - The event did NOT occur: Then \(C=0\) and \(T>X\) Eg - for patient #2, \(T > 701\)

Example - Study Data

LOW BP Group (n=7)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Number Days Followed</th>
<th>Final Status</th>
<th>(T = \text{true}) (= X)</th>
<th>((X, C))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>413</td>
<td>Dead</td>
<td>(t=413)</td>
<td>(413,1)</td>
</tr>
<tr>
<td>3</td>
<td>1075</td>
<td>Dead</td>
<td>(t=1075)</td>
<td>(1075,1)</td>
</tr>
<tr>
<td>5</td>
<td>1801</td>
<td>Dead</td>
<td>(t=1801)</td>
<td>(1801,1)</td>
</tr>
<tr>
<td>7</td>
<td>3044</td>
<td>Dead</td>
<td>(t=3044)</td>
<td>(3044,1)</td>
</tr>
<tr>
<td>8</td>
<td>3351</td>
<td>Dead</td>
<td>(t=3351)</td>
<td>(3351,1)</td>
</tr>
<tr>
<td>9</td>
<td>5551</td>
<td>Dead</td>
<td>(t=5551)</td>
<td>(5551,1)</td>
</tr>
<tr>
<td>10</td>
<td>6277</td>
<td>Dead</td>
<td>(t=6277)</td>
<td>(6277,1)</td>
</tr>
</tbody>
</table>

NORMAL BP Group (n=6)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Number Days Followed</th>
<th>Final Status</th>
<th>(T = \text{true})</th>
<th>((X, C))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>701</td>
<td>Alive</td>
<td>(t=\text{unknown})</td>
<td>(701,0)</td>
</tr>
<tr>
<td>4</td>
<td>1735</td>
<td>Dead</td>
<td>(t=1735)</td>
<td>(1735,1)</td>
</tr>
<tr>
<td>6</td>
<td>2989</td>
<td>Dead</td>
<td>(t=2989)</td>
<td>(2989,1)</td>
</tr>
<tr>
<td>11</td>
<td>7293</td>
<td>Alive</td>
<td>(t=\text{unknown})</td>
<td>(7293,0)</td>
</tr>
<tr>
<td>12</td>
<td>7352</td>
<td>Alive</td>
<td>(t=\text{unknown})</td>
<td>(7352,0)</td>
</tr>
<tr>
<td>13</td>
<td>7434</td>
<td>Alive</td>
<td>(T=\text{unknown})</td>
<td>(7434,0)</td>
</tr>
</tbody>
</table>

Thus, our data set is consists of \(n=13\) pairs \((x_1,c_1)\), \((x_2,c_2)\), \ldots, \((x_{13},c_{13})\)
**Introduction to Survival Analysis**

This is the true time to event

Previously, we worked with the ideas of a **probability distribution**, a **probability density**, and (less often, admittedly) a **cumulative distribution function**. Recall what these are about.

- For a **discrete** random variable \( X \), we defined a **probability distribution** by listing all possible outcomes together with their associated probabilities.

<table>
<thead>
<tr>
<th>Ingredient 1 - Possible value of ( X ) is represented as ( x )</th>
<th>Ingredient 2 - Probability ( \Pr[ X = x ] = f_X(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = male [ 0.53 ]</td>
<td>1 = female [ 0.47 ]</td>
</tr>
<tr>
<td>Be sure to check that this enumeration of all possible outcomes is “exhaustive”</td>
<td>Be sure to check that these probabilities add up to 100% or a total of 1.00</td>
</tr>
</tbody>
</table>

- This intuition was extended to develop a definition of a **probability density**, \( f_X(x) \), for a **continuous** random variable \( X \).

<table>
<thead>
<tr>
<th>Discrete Random Variable</th>
<th>Continuous Random Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st: “List” of all possible values that exhaust all possibilities</td>
<td>“List” ( \rightarrow ) range</td>
</tr>
<tr>
<td>E.g. – 1, 2, 3, 4, \ldots, ( \infty )</td>
<td>E.g. ( -\infty ) to ( +\infty )</td>
</tr>
</tbody>
</table>

| 2nd: **Accompanying probabilities of “each value”** | “Point probability” \( \rightarrow \) probability density |
| \( \Pr[ X = x ] \) | Probability density of \( X \), written \( f_X(x) \) |

| Total must be 1 | “Unit total” \( \rightarrow \) unit integral |
| \( \sum_{x=\text{min}}^{\max} \Pr[ X = x ] = 1 \) | \( \int_{-\infty}^{\infty} f_X(x) \, dx = 1 \) |
A little more extension gave us the notion of a cumulative distribution, $F_X(x)$, or cdf, which speaks to the question “what is the likelihood of an outcome less than or equal to ..?”

- For discrete variable:
  \[ F_X[x] = \text{Prob}[X \leq x^*] = \sum_{x=\text{lowest}}^{x^*} \text{Prob}[X=x] \]

- For continuous variable:
  \[ F_X[x^*] = \text{Prob}[X \leq x^*] = \int_{-\infty}^{x^*} f_X(x)dx \]

Thus, the probability density function $f_X(x)$ and the cumulative probability function, $F_X(x)$ are two (equivalent) ways of talking about the same thing – the distribution of a random variable. Actually, the distribution of a random variable has five representations that are equivalent, two of which are very useful to survival analysis methodology: 1) Survivor function, $S$; and 2) Hazard function, $h$.

Consider $T$: a continuous random variable describing time to event of death ($T=$time to event):

<table>
<thead>
<tr>
<th>Representation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density, $f$</strong></td>
<td>$f_T(t^*)$</td>
</tr>
<tr>
<td>“Likelihood of $T$ at $T=t^*$ ” becomes:</td>
<td></td>
</tr>
<tr>
<td>“Probability of death in a small interval of time about $t^*$”</td>
<td></td>
</tr>
<tr>
<td><strong>Cumulative distribution or “cdf”, $F$</strong></td>
<td>$F_T(t^<em>) = \text{Prob}[T&lt;t^</em>] = \int_{-\infty}^{t^*} f_T(t)dt$</td>
</tr>
<tr>
<td>“Probability $T$ is less than or equal to $t^*$”</td>
<td></td>
</tr>
<tr>
<td>“Probability individual dies before time $t^*$”</td>
<td></td>
</tr>
<tr>
<td><strong>Survival function, $S$</strong></td>
<td>$S_T(t^<em>) = \int_{t^</em>}^{\infty} f_T(t)dt = 1 - F_T(t^*)$</td>
</tr>
<tr>
<td>“Probability $T$ is greater than or equal to $t^*$”</td>
<td></td>
</tr>
<tr>
<td>“Probability individual survives beyond time $t^*$”</td>
<td></td>
</tr>
<tr>
<td><strong>Hazard function, $h$</strong></td>
<td>$h_T(t^<em>) = \lim_{\Delta t \to 0} \left( \frac{\text{Prob}[t^</em> \leq T &lt; t^* + \Delta t</td>
</tr>
<tr>
<td>“Instantaneous probability of event at time $t^<em>$ given survival up to time $t^</em>$”</td>
<td></td>
</tr>
<tr>
<td>“Probability individual dies within a small interval of time about $t^<em>$ given that he/she has survival up until time $t^</em>$”</td>
<td></td>
</tr>
</tbody>
</table>
Note: It is also possible to define a cumulative(integrated) hazard function:

- \( \Lambda_T(t) = \int_0^t h_T(t) \, dt \)

- This is useful in survival analysis, too, but is not discussed here.
Introduction to Survival Analysis

A characteristic of Survival Data is Censoring

- An observation is said to be censored if it is an incomplete observation of the actual time to event. There are three kinds of censoring:
  - **RIGHT** censored at time “t” means: “true survival time is at least t” \( (T > t) \)
  - **LEFT** censored at time “t” means: “true survival time is at most t”. \( (T < t) \)
  - **INTERVAL** censored means: “true survival is known only to have occurred somewhere within a known interval of time.

- **Unit 8 – Introduction to Survival Analysis** considers only **RIGHT** censored survival data situations. This is what you are most likely to encounter.

(Right) Censoring can Occur for a Variety of Reasons

- **The study ends.** At the time of data analysis (typically the scheduled end of study), the subject has not yet experienced the event of interest.
  - E.g. – In a study of the event of death, the study ends with some study participants still alive at trial’s end.

- **Loss to follow-up.** The subject is lost to follow-up without having experienced the event of interest.
  - E.g. – In a study of the event of death, a study participant moves and is no longer locatable.

- **Competing event.** A competing event occurs and precludes occurrence of the event of interest.
  - E.g. – In a study of event of death from cancer, a study participant dies prematurely due to a car accident.

- **Drop-out.** The subject is dropped from the study without having experienced the event of interest.
  - E.g. – In a clinical trial of event of death following treatment for HIV, a study participant is censored for non-compliance with treatment.
4. Probability Models for Survival Data

\[ T = \text{time-to-event} \]

**Introduction**

- The normal distribution is not a good model for the description of survival time random variables of the form \( T = \text{time-to-event} \). Nor are the student-t, chi square, or F distributions.

- What follows is a description of two of the more commonly used survival (also called failure time) distributions for \( T=\text{time-to-event} \): exponential and weibull. The names of some others are mentioned for reference.

- Survival time distributions are often expressed in their hazard function (h) representations, rather than in terms of other representations (such as density function, cumulative density function, or survival function).

- Graphical diagnostics to assess the appropriateness of these distributions for a given set of data are often made using the transformation \( Y = \ln [T] \).

- Survival analysis often focuses on the natural logarithm of the hazard function. Thus, the descriptions below include a description of \( \ln[ h(t) ] \).
**Exponential Distribution**

- In words: “At every point in time over the course of follow-up, the instantaneous probability of event in the next instant is the same. It doesn’t change with time”.

- “Constant hazard over time”, “memory less”, “always as good as new”.

- $h(t) = \lambda$, is constant, positive. $\lambda > 0$.

- $\ln [h(t)] = \mu$ is constant, where $\mu = \ln[\lambda]$.

- Four equivalent representations of this distribution:
  1. Hazard function: $h_t(t) = \lambda$
  2. Survival function: $S_t(t) = \text{Prob}[T > t] = \exp(-\lambda t)$
  3. Cumulative distribution function: $F_t(t) = \text{Prob}[T \leq t] = 1 - \exp(-\lambda t)$
  4. Density function: $f_t(t) = \lambda \exp(-\lambda t)$

- Diagnostic plot: Plot $\ln [ -\ln S(t) ]$ on the vertical versus $\ln (t)$ on the horizontal. The plot should approximate a straight line.

**Weibull Distribution**

- In words: “The hazard function changes with the passage over time. In particular, it follows a power function of time”.

- The $\ln [h(t)]$ is linear in $\ln (t)$.

- The hazard function, $h(t)$, is determined by two parameters, $\lambda$ and $\beta$.
  - $\lambda$ is the scale parameter
  - $\beta$ is the shape parameter

- Four equivalent representations of this distribution:
  1. Hazard function: $h_T(t) = \lambda \beta (\lambda t)^{\beta - 1}$
  2. Survival function: $S_T(t) = \text{Prob}[T > t] = \exp(-\lambda t)^\beta$
  3. Cumulative distribution function: $F_T(t) = \text{Prob} [T < t] = 1 - \exp(-\lambda t)^\beta$
  4. Density function: $f_T(t) = \lambda \beta (\lambda t)^{\beta - 1} \exp\left(-[\lambda t]^\beta\right)$
Other Commonly Used Probability Models for Survival Data

- **Gompertz Distribution:** This model assumes that $\ln \left[ h(t) \right]$ is a linear function of time. A feature of this model is that it “allows for cure”.
  
  $\ln \left[ h(t) \right] = \mu + \alpha t$

- **Piecewise Exponential:** This model also “allows for cure”. Specifically, as the name implies, the hazard is constant over intervals but can be equal to a different constant in different intervals. For example, it might be
  
  - $h(t) = \lambda$ in the interval $0 \leq t < t^*$
  - $h(t) = 0$ for $t > t^*$

- **Log-normal Distribution:** This model assumes that the natural logarithm of the time to event variable $T$ is distributed normal.
  
  $\ln \left[ (t) \right]$ is distributed Normal

- **Log-logistic Distribution:** This model is useful for describing data for which the hazard function, $h$, has a graph that looks like an inverted “U”.
  
  - $h(t)$ is low initially
  - $h(t)$ is then high
  - Eventually, $h(t)$ declines.
5. The Kaplan-Meier Curve
Model Free Estimation

Introduction

- A full treatment of survival data, including the fitting of probability models such as the ones previously discussed, is beyond the scope of this unit.

- This is an introduction that highlights the more commonly used techniques for
  1. Summarizing and graphing survival data (Kaplan-Meier Estimation)
  2. Model free comparison of two empirical survival distributions (Log rank test)
  3. Estimation of crude and adjusted hazard rates (Cox proportional hazards modeling)

The Intuition of Kaplan-Meier Estimation is the idea of “event after event after event, etc” - chronological accumulation of probabilities. The formal name for this is the theorem of total probabilities.

Imagine that your child has been given a lucky coin. He has hidden it in a drawer of one of two boxes which are located in his bedroom closet. Box 1 has 3 drawers. Box 2 has 2 drawers.

Suppose further that your child has different feelings about Box 1 and Box 2. Box 1 is covered with lots of sparkly stickers; 90% of the time, items go here. The rest of the time, 10% of the time, items go into Box 2 instead. Now suppose that, having selected a box, your child then tosses the item “willy nilly” into a drawer: at random and with equal probability.
Thus, the particulars of your lucky coin hunt are the following:

| Probability [Coin is in Box #1] | .9 90% of the time your child chooses box #1 |
| Probability [Coin is in Box #2] | .1 1.0 total |

| Given Coin is in Box #1 |
| Probability [coin is in drawer #1] | .333 |
| Probability [coin is in drawer #2] | .333 |
| Probability [coin is in drawer #3] | .333 |
| 1.000 total |
| Having chosen box #1, drawer #1 is chosen 33% of the time. Ditto for drawers #2 and #3. |

| Given Coin is in Box #2 |
| Probability [coin is in drawer #1] | .5 |
| Probability [coin is in drawer #2] | .5 |
| 1.0 total |
| Having chosen box #2, drawer #1 is chosen 50% of the time. Ditto for drawer #2. |

Suppose, now, we want to know: What is the probability that the coin is in drawer #2 of box #2?

In order for this to happen: (1) first, your child has to have selected box #2, and then (2) second, your child had to have selected drawer #2. Thus

**Answer:**

Probability [coin is in drawer #2 of box #2]

= Probability [Box #2 selected] x Probability [drawer #2 selected given Box #2 selected]

= [.1] x [.5]

= .05

- **Ah ha!** You arrived at the answer quite naturally by thinking in steps! What had to have happened first is that your child selected Box 2. The probability of this event is an example of an unconditional probability. Having accomplished this, that is “given that box #2 was selected”, the next probability (the selection of a drawer within box #2) is an example of a conditional probability.

- This is the idea of the theorem of total probabilities and, here, is just a fancy way of saying “think in steps”. This is the tool used in Kaplan-Meier Curve estimation.
**How Kaplan-Meier Curve Estimation Works**

- The chronology that is worked with is the occasions of actual events. We process them one at a time, in order.

- For each actual event time, we consider the **instant just before**:
  - By convention, we might let \( t \) denote actual event time. And \( t^- \) denote the instant just before. (note the superscript)
  - How many individuals: (1) have not yet failed; and (2) are still under observation? These persons are called “at risk”
  - Of these, how many (and %) are observed to survive beyond the occasion of the actual event time.

- Like the lucky coin example, we calculate and accumulate probabilities of the form

\[
\Pr[T>2] = \Pr[T>0] \cdot \Pr[T>1 \text{ given } T>0] \cdot \Pr[T>2 \text{ given } T>1]
\]

- The censored individuals are included in the “at risk” sets as long as it is appropriate to do so.

- **Kaplan-Meier Methodology produces model free estimates of \( \Pr[T>t] \) using observations of actual event times.**
Illustration: Case Study Data (Pooled)

It helps to construct a little worksheet:
As you read down, notice that “conditional % surviving” is updated ONLY after an actual event.

<table>
<thead>
<tr>
<th>ID</th>
<th>x</th>
<th>c</th>
<th>t of event or censoring</th>
<th># At Risk at t instant before</th>
<th># Surviving beyond t</th>
<th>Conditional % Surviving beyond t</th>
<th># At Risk to carry forward</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>413</td>
<td>1</td>
<td>413</td>
<td>13</td>
<td>12</td>
<td>12/13 = .9231</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>701</td>
<td>0</td>
<td>1 loss to censoring</td>
<td>12</td>
<td>-</td>
<td>.9231 UNCHANGED</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>1075</td>
<td>1</td>
<td>1075</td>
<td>11</td>
<td>10</td>
<td>10/11 = .9091</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1735</td>
<td>1</td>
<td>1735</td>
<td>10</td>
<td>9</td>
<td>9/10 = .90</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>1810</td>
<td>1</td>
<td>1810</td>
<td>9</td>
<td>8</td>
<td>8/9 = .8889</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>2989</td>
<td>1</td>
<td>2989</td>
<td>8</td>
<td>7</td>
<td>7/8 = .8750</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>3044</td>
<td>1</td>
<td>3044</td>
<td>7</td>
<td>6</td>
<td>6/7 = .8571</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>3351</td>
<td>1</td>
<td>3351</td>
<td>6</td>
<td>5</td>
<td>5/6 = .8333</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>5551</td>
<td>1</td>
<td>5551</td>
<td>5</td>
<td>4</td>
<td>4/5 = .80</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>6277</td>
<td>1</td>
<td>6277</td>
<td>4</td>
<td>3</td>
<td>¾ = .75</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>7293</td>
<td>0</td>
<td>1 loss to censoring</td>
<td>3</td>
<td>-</td>
<td>.75 UNCHANGED</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>7352</td>
<td>0</td>
<td>1 loss to censoring</td>
<td>2</td>
<td>-</td>
<td>.75 UNCHANGED</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>7434</td>
<td>0</td>
<td>1 loss to censoring</td>
<td>1</td>
<td>-</td>
<td>.75 UNCHANGED</td>
<td>0</td>
</tr>
</tbody>
</table>

Key: ID - Subject Identifier, X - Time on Study, C - Censoring Indicator (C=1 if Event of Death, C=0 if Censored)

Estimation of the Kaplan-Meier Curve:

<table>
<thead>
<tr>
<th>Actual Event times t</th>
<th>S[t] = Marginal Probability of Surviving Beyond t</th>
<th>S[t]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>S[0]=Pr[T&gt;0]=1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>413</td>
<td>S[413]=Pr[T&gt;413]</td>
<td>.9231</td>
</tr>
<tr>
<td></td>
<td>=Pr[T&gt;0]Pr[T&gt;413</td>
<td>T&gt;0]</td>
</tr>
<tr>
<td></td>
<td>=[1.0][.9231]</td>
<td></td>
</tr>
<tr>
<td>1075</td>
<td>S[1075]=Pr[T&gt;1075]</td>
<td>.8392</td>
</tr>
<tr>
<td></td>
<td>=Pr[T&gt;0]Pr[T&gt;413</td>
<td>T&gt;0]Pr[T&gt;1075</td>
</tr>
<tr>
<td></td>
<td>=[1.0][.9231][.9091]</td>
<td></td>
</tr>
<tr>
<td>1735</td>
<td>S[1735]=Pr[T&gt;1735]</td>
<td>.7553</td>
</tr>
<tr>
<td></td>
<td>=Pr[T&gt;0]Pr[T&gt;413</td>
<td>T&gt;0]Pr[T&gt;1075</td>
</tr>
<tr>
<td></td>
<td>=[1.0][.9231][.9091][.90]</td>
<td></td>
</tr>
<tr>
<td>1810</td>
<td>S[1810]=Pr[T&gt;1810]</td>
<td>.6714</td>
</tr>
<tr>
<td></td>
<td>=Pr[T&gt;0]Pr[T&gt;413</td>
<td>T&gt;0]Pr[T&gt;1075</td>
</tr>
<tr>
<td></td>
<td>=1.0][.9231][.9091][.90][.8889]</td>
<td></td>
</tr>
<tr>
<td>2989</td>
<td>S[2989]=Pr[T&gt;2989]</td>
<td>.5874</td>
</tr>
<tr>
<td></td>
<td>=Pr[T&gt;0]Pr[T&gt;413</td>
<td>T&gt;0]Pr[T&gt;1075</td>
</tr>
<tr>
<td></td>
<td>=1.0][.9231][.9091][.90][.8889][.8750]</td>
<td></td>
</tr>
<tr>
<td>Event times t</td>
<td>S[t] = Marginal Probability of Surviving Beyond t</td>
<td>S[t]</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3044</td>
<td>$S[3044] = Pr[T&gt;3044]$</td>
<td>.5035</td>
</tr>
<tr>
<td></td>
<td>$= Pr[T&gt;0]Pr[T&gt;413]Pr[T&gt;1075]Pr[T&gt;413]Pr[T&gt;1735]Pr[T&gt;1735]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;1810]Pr[T&gt;1735]</td>
<td>Pr[T&gt;2989]Pr[T&gt;1810]Pr[T&gt;3044]$</td>
</tr>
<tr>
<td></td>
<td>$= [1.0][.9231][.9091][.90][.8889][.8750][.8571]$</td>
<td></td>
</tr>
<tr>
<td>3351</td>
<td>$S[3351] = Pr[T&gt;3351]$</td>
<td>.4196</td>
</tr>
<tr>
<td></td>
<td>$= Pr[T&gt;0]Pr[T&gt;413]Pr[T&gt;1075]Pr[T&gt;413]Pr[T&gt;1735]Pr[T&gt;1735]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;1810]Pr[T&gt;1735]</td>
<td>Pr[T&gt;2989]Pr[T&gt;1810]Pr[T&gt;3044]$</td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;3351]$ $Pr[T&gt;3044]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$= [1.0][.9231][.9091][.90][.8889][.8750][.8571][.8333]$</td>
<td></td>
</tr>
<tr>
<td>5551</td>
<td>$S[5551] = Pr[T&gt;5551]$</td>
<td>.3357</td>
</tr>
<tr>
<td></td>
<td>$= Pr[T&gt;0]Pr[T&gt;413]Pr[T&gt;1075]Pr[T&gt;413]Pr[T&gt;1735]Pr[T&gt;1735]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;1810]Pr[T&gt;1735]</td>
<td>Pr[T&gt;2989]Pr[T&gt;1810]Pr[T&gt;3044]$</td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;3351]$ $Pr[T&gt;3044]$ $Pr[T&gt;5551]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$= [1.0][.9231][.9091][.90][.8889][.8750][.8571][.8333][.80]$</td>
<td></td>
</tr>
<tr>
<td>6277</td>
<td>$S[6277] = Pr[T&gt;6277]$</td>
<td>.2517</td>
</tr>
<tr>
<td></td>
<td>$= Pr[T&gt;0]Pr[T&gt;413]Pr[T&gt;1075]Pr[T&gt;413]Pr[T&gt;1735]Pr[T&gt;1735]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;1810]Pr[T&gt;1735]</td>
<td>Pr[T&gt;2989]Pr[T&gt;1810]Pr[T&gt;3044]$</td>
</tr>
<tr>
<td></td>
<td>$\cdot Pr[T&gt;3351]$ $Pr[T&gt;3044]$ $Pr[T&gt;5551]$ $Pr[T&gt;6277]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$= [1.0][.9231][.9091][.90][.8889][.8750][.8571][.8333][.80][.75]$</td>
<td></td>
</tr>
</tbody>
</table>

**R Illustration**

```r
library(survival)
load(file="unit8_page11.Rdata")
data.km <- survival::survfit(Surv(fu_days,dead)-1,data=data)
plot(data.km, conf.int=FALSE,
     main="Overall",
     ylab="Estimated survival probability",
     xlab="Days Since MI")
summary(data.km)
```
Features of a Kaplan-Meier Curve

- **How to** interpret this step function.
  - At time $t=0$, the estimated probability of continued survival is 1.0
  - The estimated probability of survival remains at 1.0 until the next actual event time. For these data, this is at $t=413$.
  - At time $t=413$, the estimated probability of continued survival = .92
  - Etc.

- As time progresses, the extent of data diminishes. Inclusion of Greenwood confidence intervals about the Kaplan-Meier curve (not described in this lecture) reflects this in that the width of the confidence band gets larger with time (admittedly, a little difficult to see!).

```r
Call: survfit(formula = Surv(fu_days, failure) ~ 1, data = data, conf.int = FALSE)

  time n.risk n.event survival std.err
1     413     13       1    0.923  0.0739
2   1075     11       1    0.839  0.1045
3   1735     10       1    0.755  0.1232
4   1801      9       1    0.671  0.1351
5   2989      8       1    0.587  0.1419
6   3044      7       1    0.503  0.1443
7   3351      6       1    0.420  0.1426
8   5551      5       1    0.336  0.1366
9   6277      4       1    0.252  0.1256
```

Call: survfit(formula = Surv(fu_days, failure) ~ 1, data = data, conf.int = FALSE)

- **Nature** Population/
- **Sample** Observation/
- **Data** Relationships/
- **Modeling** Analysis/
- **Synthesis**
R Illustration

```r
library(survival)

# fit Kaplan Meier model for overall data with CI
data.km.2 <- survival::survfit(Surv(fu_days, dead) ~ 1, data = data,
               error="greenwood", conf.type="log-log")

plot(data.km.2, main="Overall with 95% Greenwood CI",
ylab="Estimated survival probability",
xlab="Days since MI")

summary(data.km.2)
```

<table>
<thead>
<tr>
<th>time</th>
<th>n.risk</th>
<th>n.event</th>
<th>survival</th>
<th>std.err</th>
<th>lower 95% CI</th>
<th>upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>413</td>
<td>13</td>
<td>1</td>
<td>0.923</td>
<td>0.0739</td>
<td>0.5664</td>
<td>0.989</td>
</tr>
<tr>
<td>1075</td>
<td>11</td>
<td>1</td>
<td>0.839</td>
<td>0.1045</td>
<td>0.4940</td>
<td>0.957</td>
</tr>
<tr>
<td>1735</td>
<td>10</td>
<td>1</td>
<td>0.755</td>
<td>0.1232</td>
<td>0.4161</td>
<td>0.914</td>
</tr>
<tr>
<td>1801</td>
<td>9</td>
<td>1</td>
<td>0.671</td>
<td>0.1351</td>
<td>0.3422</td>
<td>0.862</td>
</tr>
<tr>
<td>2989</td>
<td>8</td>
<td>1</td>
<td>0.587</td>
<td>0.1419</td>
<td>0.2738</td>
<td>0.804</td>
</tr>
<tr>
<td>3044</td>
<td>7</td>
<td>1</td>
<td>0.503</td>
<td>0.1443</td>
<td>0.2110</td>
<td>0.739</td>
</tr>
<tr>
<td>3351</td>
<td>6</td>
<td>1</td>
<td>0.420</td>
<td>0.1426</td>
<td>0.1541</td>
<td>0.668</td>
</tr>
<tr>
<td>5551</td>
<td>5</td>
<td>1</td>
<td>0.336</td>
<td>0.1366</td>
<td>0.1837</td>
<td>0.591</td>
</tr>
<tr>
<td>6277</td>
<td>4</td>
<td>1</td>
<td>0.252</td>
<td>0.1256</td>
<td>0.0607</td>
<td>0.507</td>
</tr>
</tbody>
</table>
Here’s a comparison of the Kaplan-Meier Curves for the Same Data, broken down by group: LOW BP versus NORMAL BP

R Illustration

```r
library(survival)

# fit Kaplan Meier model by group
data.km.3 <- survival::survfit(Surv(fu_days, dead) ~ group, data = data, conf.int = FALSE)

plot(data.km.3, lty = 1:2, main = "Kaplan Meier estimated survival by group", 
     col = c("darkblue", "darkred"), 
     ylab = "Estimated survival probability", 
     xlab = "Days since MI")
legend("topright", legend = c("Normal BP","Low BP"), 
       lty = 1:2, col = c("darkblue", "darkred"))

summary(data.km.3)
```

Call: survfit(formula = Surv(fu_days, failure) ~ group, data = data, 
              conf.int = FALSE)

```r
  group=Normal BP
     time n.risk n.event survival std.err
1735      5       1      0.8   0.179  
2989      4       1      0.6   0.219  

  group=Low BP
     time n.risk n.event survival std.err
  413      7       1  0.857   0.132  
1075      6       1  0.714   0.171  
1801      5       1  0.571   0.187  
3044      4       1  0.429   0.187  
3351      3       1  0.286   0.171  
5551      2       1  0.143   0.132  
6277      1       1  0.000  NaN
```

- Among the 13 individuals included in this case study, it appears that normal BP at the time of MI is associated with a better prognosis.
- In the next section, we learn the Log Rank test. This is a tool for assessing the statistical significance of the departure of the two survival patterns from one another.
6. The Log Rank and Related Tests

Model Free Comparison

The Log Rank Test

For two independent groups with survival distributions $S_1(t)$ and $S_2(t)$, respectively:

Null Hypothesis, $H_0$: $S_1(t) = S_2(t)$  “Equality” of survival over time
Alternative Hypothesis, $H_A$: $S_1(t) \neq S_2(t)$

Introduction to the Log Rank Test

- **Remember the analysis of a series of K 2x2 tables?** See BIOSTATS 640 Unit 4 (*Categorical Data Analysis*). As an example, we might have had a separate 2x2 table for each of several groupings of age. There we learned the Mantel-Haenszel test for testing the null hypothesis of a unit odds ratio for a sequence of several 2x2 tables under the assumption of a constant odds ratio. The idea of the log-rank test is an extension of the Mantel-Haenszel test described for K 2x2 tables. In fact, the log-rank test is actually a Mantel-Haenszel test.

- The log rank test is a model-free statistical hypothesis test procedure for the comparison of two or more survival curves. In the following example, we compare two groups only.

- Other names for the log rank test are the Peto and Peto log-rank test and Mantel’s log-rank test.

- The appropriateness of the log rank test rests on 3 assumptions.
  1. **Independence** - The two samples are random samples from two independent groups.
  2. **Equal Censoring** - The pattern of censoring is the same in the two groups.
  3. **Proportional Hazards** - The hazard of death in group #2 is a constant multiple $\theta$ (*it’s called the hazard ratio and its interpretation is very similar to that of a relative risk*) of the hazard of death in group #1 over all occasions of time. This is known as the assumption of proportional hazards:

$$h_{\text{Group } #2}(t) = \theta [h_{\text{Group } #1}(t)] \quad \text{for all } t.$$

$$S_{\text{Group } #2}(t) = [S_{\text{Group } #1}(t)]^{\theta} \quad \text{for all } t.$$
Idea of the Log Rank Test

- Under the assumption of proportional hazards, the null hypothesis of equivalence of the two survival curves is \( H_0 : \theta = 1 \). \textit{Note – On the previous page, we noted that the validity of the log rank test requires that } \( \theta = \) ”hazard ratio” \textit{is assumed to be constant over time. The null hypothesis says that } \( \theta = 1 \), \textit{meaning that the survival distribution is the same in the two groups}.

- When the null hypothesis is true, at every occasion of death(s), the distribution of the deaths in both groups is the same.

\textbf{Example} – We continue with the case study data introduced on page 11. Is survival time after a heart attack (MI) different for persons with low versus normal BP at the time of their heart attack?

\textbf{Step 1: } Consider the actual times of death. Order these from earliest to latest.

\textbf{Example: }
The 9 actual deaths occurred at times (ordered); 413, 1075, 1735, 1810, 2989, 3044, 3351, 5551, 6277

\textbf{Step 2: } At each occasion of actual death \( t \), consider the instant just before, \( t^- \), Construct the following \textit{“observed”} 2x2 table:

<table>
<thead>
<tr>
<th></th>
<th>Dies</th>
<th>Survives</th>
<th>At Risk at ( t^- ) (the instant just before)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1: Low</td>
<td></td>
<td>By subtraction</td>
<td></td>
</tr>
<tr>
<td>Group 2: Normal</td>
<td>By subtraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( O_{1t} )</td>
<td></td>
<td>By subtraction</td>
<td></td>
</tr>
<tr>
<td>( d_t )</td>
<td></td>
<td>( (N_t - d_t) )</td>
<td></td>
</tr>
<tr>
<td>( n_{1t} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n_{2t} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( N_t )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( O_{1t} = \) # deaths in group #1 at time \( t \)
\( n_{1t} = \) # at risk in group #1 at time \( t^- \)
\( n_{2t} = \) # at risk in group #2 at time \( t^- \)
\( d_t = \) total # deaths at time \( t \)
\( N_t = \) total # at risk at time \( t^- \)

\textbf{Key:}

For example, at \( t=413 \) we have:

\textbf{Group 1: Low} \hspace{1cm} \textbf{Group 2: Normal} \\
\begin{tabular}{ccc}
\hline
Dies & Survives & At Risk at \( t=413^- \) \\
\hline
\( O_{1t} = 1 \) & 6 & \( n_{1t} = 7 \) \\
0 & 6 & \( n_{2t} = 6 \) \\
\hline
\( d_t = 1 \) & 12 & \( N_t = 13 \) \\
\hline
\end{tabular}
Step 3: **Apply the null hypothesis model to obtain the null hypothesis common probability of event in each group.** Do this at each actual occasion of event.

Under the null hypothesis of equal survival distributions in the two groups, at each actual occasion of death, the (conditional) probability of death should be the same in both groups:

The **null hypothesis common probability of event** is estimated using:

\[ \hat{p}_{\text{NULL}}(\tau) = \frac{d_t}{N_t} \]

is the overall proportion of deaths at time \( t \):

<table>
<thead>
<tr>
<th>Expected # deaths</th>
<th>Expected # surviving</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_t )</td>
<td>( N_t - d_t )</td>
</tr>
<tr>
<td>( N_t )</td>
<td></td>
</tr>
</tbody>
</table>

For example, at \( t=413 \), 1 death occurred among the 13 who are at risk at the instant before.

Thus, the overall proportion of deaths at time \( t=413 \)

\[ \hat{p}_{\text{NULL}}(\tau) = \frac{d_t}{N_t} = \frac{1}{13}. \]

Step 4: **Obtain the null hypothesis expected counts.**

Apply \( \hat{p}_{\text{NULL}} \) to each group to obtain

**Expected # = (# in group) \times \hat{p}_{\text{NULL}}.**

Thus, at \( t=413 \),

In group #1 (low blood pressure), \( n_{1t} = 7 \rightarrow E_{1t} = (7)(1/13) = 0.5385 \)

Good news – all the other null hypothesis expected counts are obtained by subtraction.

<table>
<thead>
<tr>
<th>Dies</th>
<th>Survives</th>
<th>At Risk at ( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_{1t} )</td>
<td>By subtraction</td>
<td>( n_{1t} )</td>
</tr>
<tr>
<td>( E_{2t} )</td>
<td>By subtraction</td>
<td>( n_{2t} )</td>
</tr>
<tr>
<td>( d_t )</td>
<td>( (N_t - d_t) )</td>
<td>( N_t )</td>
</tr>
</tbody>
</table>

\( E_{1t} = \frac{(1/13)[7]}{\text{Expected # deaths}} = 0.5385 \)

\( E_{2t} = \frac{(1 - 0.5385)}{\text{Expected # deaths}} = 0.4615 \)

\( E_{1t} = (7 - 0.5385) = 6.4615 \)

\( E_{2t} = (12 - 0.4615) = 5.5385 \)

\( 7 \)

\( 6 \)

\( 1 \)

\( 12 \)

\( 13 \)

\( E_{1t} = \frac{\text{Expected count}}{\text{row 1, column 1}} = (n_{1t})[\hat{p}_{\text{NULL}}(\tau)] = (n_{1t})[d_t/N_t] \)

\[ \text{Variance}[O_{1t}]_{\text{NULL TRUE}} = \text{Var}[O_{1t}] = (n_{1t}) \left[ \frac{d_t}{N_t} \right] \left[ \frac{n_{2t}(N_t - d_t)}{N_t(N_t - 1)} \right] \]

These are from the central hypergeometric.
Worksheet to illustrate hand calculation

<table>
<thead>
<tr>
<th>t</th>
<th>O₁ₜ</th>
<th>n₁ₜ</th>
<th>d₁</th>
<th>(N₁ - d₁)</th>
<th>n₂ₜ</th>
<th>N₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>413</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>12</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>1075</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>1735</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>1810</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2989</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>3044</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>3351</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>5551</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6277</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Worksheet - continued

<table>
<thead>
<tr>
<th>t</th>
<th>O₁ₜ</th>
<th>E₁ₜ = (n₁ₜ)d₁ / N₁</th>
<th>V(O₁ₜ) = [E₁ₜ] [n₂ₜ(N₁ - d₁)] / N₁(N₁ - 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>413</td>
<td>1</td>
<td>.5385</td>
<td>.2485</td>
</tr>
<tr>
<td>1075</td>
<td>1</td>
<td>.5455</td>
<td>.2480</td>
</tr>
<tr>
<td>1735</td>
<td>0</td>
<td>.5000</td>
<td>.2500</td>
</tr>
<tr>
<td>1810</td>
<td>1</td>
<td>.5556</td>
<td>.2469</td>
</tr>
<tr>
<td>2989</td>
<td>0</td>
<td>.5000</td>
<td>.2500</td>
</tr>
<tr>
<td>3044</td>
<td>1</td>
<td>.5714</td>
<td>.2449</td>
</tr>
<tr>
<td>3351</td>
<td>1</td>
<td>.5000</td>
<td>.2500</td>
</tr>
<tr>
<td>5551</td>
<td>1</td>
<td>.4000</td>
<td>.2400</td>
</tr>
<tr>
<td>6277</td>
<td>1</td>
<td>.2500</td>
<td>.1875</td>
</tr>
<tr>
<td>Totals</td>
<td>7</td>
<td>4.3610</td>
<td>2.1658</td>
</tr>
</tbody>
</table>

Log Rank Test Statistic

H₀: S₁(t) = S₂(t)
Hₐ: S₁(t) ≠ S₂(t)

\[
\chi^2_{\text{LOG RANK; 1df}} = \left( \frac{\sum_{t=1}^{#\text{deaths}} O_{1t} - \sum_{t=1}^{#\text{deaths}} E(O_{1t})}{\sum_{t=1}^{#\text{deaths}} V(O_{1t})} \right)^2
\]

Under the null, this is distributed chi square with df=1
→ Rejection occurs for large values of this test statistic
Example - continued:

\[ X_{\text{LOG RANK};1\text{df}}^2 = \left( \sum_{i=1}^{#\text{deaths}} O_{it} - \sum_{i=1}^{#\text{deaths}} E(O_{it}) \right)^2 / \sum_{i=1}^{#\text{deaths}} V(O_{it}) \]

\[ = \left( 7 - 4.3610 \right)^2 / 2.1658 = 3.22 \]

p-value = Probability [ Chi square \( \text{DF}=1 \) \( \geq 3.22 \) ] = .07

Do not reject the null. Conclude that the disparity in the two survival curves is not statistically significant.  
*Note: This is not surprising. With a sample size of 13 and only 9 events spread over two groups, statistical significance of any apparent group difference was going to be hard to attain!*

**R Illustration**

```r
library(survival)
survdiff(Surv(fu_days, dead) ~ group, data = data)
```

```
Call: survdiff(formula = Surv(fu_days, failure) ~ group, data = data)

    N Observed Expected (O-E)^2/E  (O-E)^2/V
group=Normal BP 6 2   4.64     1.5   3.22
group=Low BP   7 7   4.36     1.6   3.22

  Chisq= 3.2 on 1 degrees of freedom, p= 0.0729
```

**Graphical Check of Assumption #3 (Proportional Hazards)**

**Step 1:** For each group, separately, obtain the Kaplan-Meier estimates of survival, \( S(t) \), over time:
- For Group #1, these will be \( S_1(t) \)
- For Group #2, these will be \( S_2(t) \)

**Step 2:** Do an overlay plot of two graphs
- For Group #1, \( Y=\ln\left[ -\ln(S_1(t)) \right] \) versus \( X=\ln(t) \)
- For Group #2, \( Y=\ln\left[ -\ln(S_2(t)) \right] \) versus \( X=\ln(t) \)

**Step 3:** Look for parallel lines. The assumption of proportional hazards is satisfied when the difference between the two figures is constant over time (parallelism)
Illustration in R

```r
# Create separate data frames for "Normal" and "Low BP"
data.normal <- subset(data, group==0)
data.low <- subset(data, group==1)

# Get Kaplan Meier Curvesnormal.KM <- survfit(Surv(fu_days,dead)~1,data=data.normal)low.KM <- survfit(Surv(fu_days,dead)~1,data=data.low)

# For plotting, get Y = ln [-ln (S[t])]normalY <- log(-log(summary(normal.KM)$surv))lowY <- log(-log(summary(low.KM)$surv))

# For plotting, get X = ln [ t ]normalX <- log(summary(normal.KM)$time)lowX <- log(summary(low.KM)$time)

# Put together the Y and X for plotting into dataframesnormaldf <- data.frame(normalY, normalX)lowdf <- data.frame(lowY, lowX)

# Plot (admittedly the code is complicated)plot(lowdf$lowX, y=lowdf$lowY,lwd=2,col="blue",xlab="Time", ylab="ln(-ln(S(t)))", main="Check of PH Assumption")points(normaldf$normalX, y=normaldf$normalY, lwd=2,col="red")lines(normaldf$normalX, y=normaldf$normalY, lwd=2,col="red")lines(lowdf$lowX, y=lowdf$lowY, lwd=2,col="blue")legend("topright", legend=c("Normal", "Low"), col=c("red", "blue"), lty=1:2, cex=0.8, box.lty=0)
```

Sample sizes of 7 and 6 are too small to permit meaningful interpretation here
Overview of Related Tests

Other, related, tests are available. They differ in the extent to which weights are applied to the data.

Log Rank Test -
This statistic assigns equal weight to each observation.

Gehan Test -
This statistic assigns greater weight to the earlier observations. Its formulation derives from a procedure known as the Wilcoxon rank sum test for the nonparametric comparison of two groups (See BIOSTATS 640 Unit 3, Nonparametrics)

Tarone and Ware Test -
This statistic, in its weighting scheme, is intermediate between the Log Rank and the Gehan tests.

Breslow Test -
This statistic is a generalization of the Gehan test. It is appropriate when the number of groups being compared is greater than 2.

It is beyond the scope of this course to give a full development of each.
7. Introduction to the Cox Proportional Hazards (PH) Model

Introduction

- **The Cox proportional hazards model is a relative risk regression model.** The dependent variable T is “failure time” or “time to event”.

- Generally, the observed “time to event” data are subject to **censoring**.

- Thus, the techniques of Cox proportional hazards model development is about modeling time to event and its relationship to a set of one or more explanatory variables in the presence of censoring.

- **A characteristic feature of the Cox PH model is that it focuses on the hazard function.** This is unlike the models we discussed in units 5 (Regression and Correlation) and 7 (Logistic Regression) where the focus was on the likelihood.

Definition of the Cox Proportional Hazards (PH) Model

\[ h(t; X_1, ..., X_p) = h_0(t) \exp[\beta_1 X_1 + ... + \beta_p X_p] \]

Hazard at time=t for an individual with explanatory variables \(X_1, X_2, ..., X_p\)  

Effect of the explanatory variables \(X_1, X_2, ..., X_p\) is to act multiplicatively on the hazard. Note! There is no intercept here. This is because the "baseline hazard" is modeled in \(h_0(t)\)

Example -
In the example that we have been using (survival following a heart attack), there is just one predictor called group (1=low BP, 0=normal BP). Thus, here, \(X=\text{group}\).
Some Features of the Cox PH Model

- The ratio of hazards (relative hazard) is called a relative risk.

- A characteristic feature of the model is that the ratio of hazards (the relative risk) for any two individuals does **NOT** depend on time. (unless we specifically modify the model definition to incorporate time dependence – *more on this later*).

- Thus, we can explore the influence of multiple explanatory variables on the relative hazard through estimation of and tests of hypotheses about the regression parameters $\beta_1, \beta_2 \ldots \beta_p$.

  *Note – This is analogous to what we did in logistic regression. There, we obtained estimates of the odds ratio measure of association via OR=$\exp(\beta)$.

- Because the baseline hazard, $h_0$, is allowed to be arbitrary, we can allow it to vary across strata! Specifying a separate stratum-specific baseline hazard for each stratum of an hypothesized explanatory variable is useful when that explanatory variable does not act multiplicatively on the hazard.

  *Beware - If you do this, the explanatory variable that defines the strata CANNOT be among the predictors $X_1, \ldots, X_p$.

### Cox Proportional Hazards (PH) Model for Stratified Data

$$h_j(t; X_1, \ldots, X_p) = h_{0j}(t) \exp[ \beta_1 X_1 + \ldots + \beta_p X_p ]$$

**Hazard at time**=$t$ **for an individual in stratum** “$j$” **with explanatory variables** $X_1, X_2, \ldots, X_p$

**Multiplicative effect of the explanatory variables** $X_1, X_2, \ldots, X_p$ **is assumed to be the same across all strata** “$j$”.

**Stratum** “$j$” **baseline hazard At time**=$t$. 

| Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis |
8. Interpretation of a Cox PH Model

The Cox PH model focuses on hazard ratios (or relative risks) in its assessment of the influences of the hypothesized correlates of survival time and tests of hazard ratios = 1 become tests of \( \beta = 0 \).

- To see this, consider a one predictor model with predictor \( X \) defined as a 0/1 indicator of exposure:
  \[
  X = 1 \quad \text{if exposed (smokers)} \\
  0 \quad \text{if not exposed (Non-smokers)}
  \]

- Thus, the model of hazard at time \( t \), according to the Cox PH approach is defined as follows.
  \[
  h(t; X) = h_0 \, exp[ \beta_1 X ]
  \]

- Relative Risk at time \( t = \)
  \[
  \frac{\text{Hazard of event for } X=1}{\text{Hazard of event for } X=0} \\
  = \frac{h_0 \, exp[ \beta_1 \cdot 1]}{h_0 \, exp[ \beta_1 \cdot 0]} \\
  = \frac{h_0 \, exp[ \beta_1]}{h_0 \, [1]} \\
  = exp[\beta_1]
  \]

- Notice the reliance in this proof on the assumption of proportional hazards.
  - Therefore, appropriate use of the Cox PH model analysis approach requires an assessment of the proportional hazards assumption.

- Next we consider the interpretation of some already fitted Cox PH models.
Example – continued

Recall that our data set is comprised of 13 patients hospitalized for heart attack and interest is in the role of low blood pressure at the time of MI as a predictor of early death. Recall also that there are two groups:

Group #1: **Low Blood** Pressure Group

\[ N_1 = 7 \]
\[ n_1 = 7 \text{ events of death at times: } 413, 1075, 1810, 3044, 3351, 5551 \text{ and } 6277 \text{ days.} \]

Group #2: **Normal Blood** pressure Group

\[ N_2 = 6 \]
\[ n_2 = 2 \text{ events of death at times: } 1735 \text{ and } 1989 \text{ days.} \]

- Using *STATA*, we fit a one predictor COX PH model with predictor \( X=\text{group} \) defined as a 0/1 indicator of **LOW** blood pressure at the time of MI. Thus,

\[
X = \text{group} = \begin{cases} 
1 & \text{if patient has low BP at the time of MI} \\
0 & \text{if patient has normal BP at the time of MI} 
\end{cases}
\]

- The model fitted is the simple one we just saw:

\[
h(t; \text{group}) = h_0(t) \exp[ \beta*\text{group} ]
\]

- *R* obtains an estimated \( \hat{\beta} = 1.369 \) with associated \( se = 0.817 \)

**R Illustration**

```r
library(survival)
survival::coxph(Surv(fu_days, dead) ~ group, data = data)
```

```
Call:
  coxph(formula = Surv(fu_days, failure) ~ group, data = data)

   coef  exp(coef)   se(coef)    z      p
groupLow BP 1.369  3.933     0.817 1.68  0.094

Likelihood ratio test= 3.39  on 1 df, p=0.0654
n= 13, number of events= 9
```

| Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis |
R also produces the corresponding estimated hazard $\hat{h}=3.93$ with associated 95% CI=(0.7922, 19.52)

**Key** -

- The output you see has some other pieces of information; these pertain to model estimation which is introduced later.

- In words $\hat{h}=3.93$ says: “at all times following MI, under the assumption of proportional hazards, persons with low blood pressure at the time of their MI have a hazard (risk) of death 3.93 times greater than that of persons with normal blood pressure at the time of their MI”.

  **This is quite a lot to say, isn’t it! Remember – this statement presumes that it is okay to assume proportional hazards.**

- The output, in particular the confidence intervals, are wide and include the null hypothesis of equivalence. Thus, the apparent increased hazard of death for persons with low blood pressure is not statistically significant.

  **Wide confidence intervals are not surprising given that the total sample size is only 13, the group sample sizes are 7 and 6 and the event counts are 7 and 2. We don’t have a lot to go on here.**
Let’s try another example.
Example #2 (modified from Hand et al, 1994)

This example is a study of time in methadone treatment. Two predictors are of interest: (1) \( X_1 = \text{dose} \) of methadone, and (2) \( X_2 = \text{prison} \) defined as a 0/1 indicator of a history of incarceration. It is of interest to investigate whether dose of methadone plays a role in retention in treatment and, additionally, whether persons in methadone treatment for heroin addiction are more likely to drop out of treatment if they have a history of incarceration. A third variable, clinic, indexes the 2 methadone clinics providing data. Clinic is a stratification variable, with possibly distinct baseline hazards.

- \( T = \) “time to event” is referring to “time in methadone treatment” and “event” occurs when a resident drops out of treatment.

- The analyst allows for the possibility that the two clinics may have hazard functions that are not parallel. Thus, the following stratified Cox PH model is fit:

\[
h_j(t; X_1, X_2) = h_{j0}(t) \exp[\beta_1 X_1 + \beta_2 X_2]
\]

(1) “\( j \)” indexes stratum and refers to clinic #1 or #2

(2) \( X_1 \) is dose and is a continuous measure of methadone dose (mg)

(3) \( X_2 \) is a 0/1 indicator of history of incarceration with 1=incarcerated and 0 = not

- The following results are obtained:

|          | Haz. Ratio | Std. Error | \( z \)  | \( P > |z| \) | 95% CI      |
|----------|------------|------------|--------|-------------|------------|
| \( X_1 \) = dose | 0.97       | .006       | -5.436 | 0.000       | (.95, 0.98) |
| \( X_2 \) = prison | 1.48       | .249       | 2.302  | 0.021       | (1.06, 2.05)|

Key

- **For \( X_1 \) = dose:** Controlling for history of incarceration, it is estimated that for every unit increase in methadone dose (1 mg), the hazard of ending methadone treatment (“dropping out”) is multiplied by 0.97 (meaning they are less likely to end treatment).

- **For \( X_2 \) = prison:** Controlling for dose, it is estimated that persons with a prison history are 48% more likely to drop out of methadone treatment.
9. Hypothesis Testing Using the Cox PH Model

Global - “Omnibus” Tests that All of the \( \beta \)’s are Zero

- **Global Question:** “Is there any explanatory information in any of the \( X_1, X_2, \ldots, X_p \) under the assumption of the appropriateness of the Cox PH model formulation?” Note the reliance on the assumption of the correctness of the PH model and its assumptions!

- **H_0:** \( \beta_1=0, \beta_2=0, \ldots, \beta_p=0 \)
- **H_A:** At least one of the \( \beta_i \) is NOT 0

Recall the Overall F Test in Normal Theory Regression. In normal theory linear regression, the overall F-test was introduced as a global test of the fitted linear model, relative to the “intercept-only” model. The null hypothesis was that all of the \( \beta \)’s are zero. Rejection of the null hypothesis suggested that the fitted model explains statistically significantly more of the variability in the outcome \( Y \), than the “intercept-only” model. Reminder – rejection of the null hypothesis communicates only limited information. We still don’t know if the fitted model is a good one.

In Cox PH modeling, there are 3 tests (asymptotically equivalent) of the global null hypothesis that all of the \( \beta \)’s are zero.

- **#1. (Partial) Likelihood Ratio LR Chi Square Test:**
  - Often denoted \( G \)
  - \( G = (-2) \left[ \ln \text{-partial likelihood (null model)} - \ln \text{-partial likelihood (current model)} \right] \)
  - Under the null hypothesis, \( G \) is distributed Chi Square with \( \text{df} = \# \text{ predictors} \)
  - Rejection occurs for large values of the chi square statistic (accompanying \( p \)-values are small)

- **#2. Wald Normal(0,1) Test:**
  - This is a Z-test that can be examined for each predictor one at a time.
  - Wald Z-statistic = \( \frac{\text{estimated } \beta - 0}{\text{SE(estimated } \beta)} \)
  - Under the null hypothesis, the Wald Z-statistic is distributed Normal(0,1)
  - Rejection occurs for large values of the Wald Z-test statistic (accompanying \( p \)-values are small)

- **#3. Score Test:**
  - Often denoted \( Z^* \)
  - This is another Z-statistic test of the null hypothesis
  - Under the null hypothesis, the Score Test Z-statistic is distributed Normal(0,1)
  - Rejection occurs for large values of the Score Z-test statistic (\( p \)-values are small)

<table>
<thead>
<tr>
<th>Nature</th>
<th>Population/Sample</th>
<th>Observation/Data</th>
<th>Relationships/Modeling</th>
<th>Analysis/Synthesis</th>
</tr>
</thead>
</table>

If the Partial-LR, Z and \( Z^* \) do not agree, the Partial –LR Test should be used.
Model Building - Likelihood Ratio Test Comparison of “Hierarchical” Models

The likelihood ratio test in Cox PH model analyses of survival data is analogous to the likelihood ratio test in logistic regression analyses. Two “hierarchical” models are being compared. The “null” model is a “reduced” or “smaller” model. The “alternative” model is the “enhanced” model. It contains all the predictors that are in the null model plus some extra predictors. Are the “extra” predictors important, after controlling for all the variables in the smaller model?

- **Model Comparison Question:** “Is there any additional, statistically significant, explanatory information in any of the extra predictors \( X_{p+1}, X_{p+2}, \ldots, X_{p+k} \) after controlling for \( X_1, X_2, \ldots, X_p \)?

- **\( H_0 \):** Controlling for \( X_1, X_2, \ldots, X_p, X_{p+1}, X_{p+1}, \ldots, X_{p+k} \) are not statistically significant \( \Rightarrow \beta_{p+1}=0, \beta_{p+2}=0, \ldots, \beta_{p+k}=0 \)

- **\( H_A \):** At least one of the \( \beta_{p+i} \) is NOT 0

- The likelihood ratio test here is actually a comparison of two “\(-2 \ln \) partial likelihoods”. If you are interested, see the “Appendix” for an introduction to partial likelihood.

<table>
<thead>
<tr>
<th>Likelihood ratio Test (LR Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison of “Hierarchical Models”</td>
</tr>
</tbody>
</table>

\[
H_0: \text{Controlling for } X_1, X_2, \ldots, X_p, X_{p+1}, X_{p+1}, \ldots, X_{p+k} \text{ are not statistically significant } \Rightarrow \beta_{p+1}=0, \beta_{p+2}=0, \ldots, \beta_{p+k}=0
\]

\[
H_A: \text{At least one of the } \beta_{p+i} \text{ is NOT } 0
\]

LR Test (df = k) \( = -2 \ln \left[ \frac{\text{partial likelihood (data using } \hat{\beta}_{\text{MLE; null}} \text{ )}}{\text{partial likelihood (data using } \hat{\beta}_{\text{MLE; alternative}} \text{ )}} \right] \)

Under the null, this is distributed chi square with df=k (the # extra predictors)

\( \Rightarrow \) Rejection occurs for large values of this test statistic
10. Evaluating the Proportional Hazards Assumption

- In the introduction so far, it has been noted and emphasized that the proportional hazards assumption is at the heart of the Cox PH model.

- Recall what this assumption is saying: “At every point in the time range under study, the hazard of event in the comparison group is a constant multiple of the hazard of event in the reference group”

- Example 1 - continued: We estimated that “at all times following MI, persons with low blood pressure at the time of their MI have a hazard of death 3.93 times greater than that of persons with normal blood pressure at the time of their MI”…. under the assumption of proportional hazards...

- An overview of some of the more commonly used techniques for assessing the proportional hazards assumption are presented.

Graphical Assessment #1 - Log Log Plots

- Previously, we learned that a graphical assessment is \( Y = \ln[\ln(S[t])] \) versus \( X = \text{time} \) and that, under proportional hazards, these graphs will show a constant difference on the vertical (a kind of parallelism). *See again, page 35.*

- Another plot that might be easier to look at is an overlay plot of \( Y = \ln[\ln(S[t])] \) versus \( X = \ln[\text{time}] \) for each group defined by the predictor variable

- Under proportional hazards, the overlaid plots should be straight lines that are parallel.

- **R command**
  
  ```
  mod <- coxph(Surv(fu_days, dead) ~ group, data = data)
  plot(survfit(mod))
  ```
Graphical Assessment #2 - Comparison of Cox Model w Kaplan Meier

- Recall that the Kaplan-Meier methodology for estimating a survival curve is model free. In contrast the Cox PH model presumes the proportional hazards assumption.

- The idea of this plot is to compare estimates of survival obtained by Kaplan Meier methodology with estimates of survival obtained via estimation of a Cox PH model.
  - Obtain Kaplan Meier curve for each value of predictor.
  - Fit Cox PH model using the same predictor as the explanatory variable.

R commands

Forthcoming ...

Here is an illustration (taken from the same lecture by Dr. Kerr):

KEY: In this graph, “Observed” denotes the Kaplan-Meier curve while the “Predicted” denotes the fit of the Cox PH model.
Hypothesis Test of the Proportional Hazards Assumption

- A statistical hypothesis test of the proportional hazards assumption is obtained by including a special extra explanatory variable in the model and assessing the departure of its regression coefficient from zero.

- Example 1, continued- If we suspect that the proportional hazards assumption is not satisfied for the predictor variable indicating blood pressure at the time of MI (X=1 if low, X=0 if normal), then we might fit a model that contains the additional predictor defined:

\[ X_{\text{time}} = [X] \times [\ln(t)] \]

- Rejection of the null hypothesis that the regression parameter \( \beta \) is zero is evidence of a violation of the assumption of proportional hazards.
11. Regression Diagnostics for the Cox PH Model

Yes, we need regression diagnostics here, too! Just as we learned for assessing the fitted model in normal theory linear regression and in logistic regression, regression diagnostics are of two types, systematic and case specific. The following is a partial listing of the regression diagnostics questions we might ask:

- **Systematic component**
  
  o Is the fit of the Cox PH model a reasonably good fit to the observed data?
  
  o Is the assumption of proportional hazards reasonable?

- **Case analysis**
  
  o Is the fitted model excessively influenced by one or a small number of individuals?

**For R Users:**

There exist methods to address each of these regression diagnostic questions. As these are all “post-estimation” commands, you must have fit a model first.

<table>
<thead>
<tr>
<th>Question</th>
<th>Method of Assessment</th>
</tr>
</thead>
</table>
| Is the fit of the Cox PH model a reasonably good fit to the observed data? | mod <- coxph(Surv(y, ind) ~ ., data = data)
summary(mod)

**What to look for:** NON statistical significance suggesting that the fit is a reasonably good fit. |
| Is the fit of the Cox PH model a reasonably good fit to the observed data? | mod1 <- survfit(Surv(y, ind) ~ 1, data = data, conf.int = FALSE)
mod2 <- coxph(Surv(y, ind) ~ ., data = data)
fit <- survfit(mod2)
plot(fit$surv, fit$time, type = "s")
lines(mod1$surv, mod1$time, col = "red", type = "s")

**What to look for:** This is a graphical assessment of how well the COX PH model fits the data. Observed=Kaplan-Meier, Predicted = COX PH fit. Look for similarity of observed and predicted. |
| Is the assumption of proportional hazards met? | mod <- coxph(Surv(y, ind) ~ ., data = data)
cox.zph(mod)

**What to look for:** The function cox.zph() [in the survival package] provides a convenient solution to test the proportional hazards assumption for each covariate included in a Cox regression model fit. For each covariate, the function cox.zph() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole. |
<table>
<thead>
<tr>
<th>Question</th>
<th>Method of Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the assumption of proportional hazards met?</td>
<td><code>mod &lt;- coxph(Surv(y, ind) ~ ., data = data)</code></td>
</tr>
<tr>
<td></td>
<td><code>summary(mod)</code></td>
</tr>
<tr>
<td></td>
<td><strong>What to look for:</strong> This is a formal test of the proportional hazards assumption (null: proportional hazards assumption is met). Look for NON statistical significance.</td>
</tr>
<tr>
<td>Is the assumption of proportional hazards met?</td>
<td><code>mod &lt;- coxph(Surv(y, ind) ~ ., data = data)</code></td>
</tr>
<tr>
<td></td>
<td><code>plot(mod$residuals)</code></td>
</tr>
<tr>
<td></td>
<td><strong>What to look for:</strong> This is a graphical assessment of proportional hazards. Look for trend of the residuals of this fit.</td>
</tr>
<tr>
<td>Is the functional form for a particular predictor okay or is some sort of transformation needed?</td>
<td><code>.lowess residualsnamel predictorname, mean</code></td>
</tr>
<tr>
<td></td>
<td><strong>What to look for:</strong> A smooth and flat plot suggests that all is well. A curved plot suggests the need for a transformation.</td>
</tr>
<tr>
<td>Is the proportional hazards assumption okay for a particular predictor?</td>
<td><code>mod &lt;- coxph(Surv(y, ind) ~ ., data = data)</code></td>
</tr>
<tr>
<td></td>
<td><code>plot(mod$residuals,mod$time)</code></td>
</tr>
<tr>
<td></td>
<td><strong>What to look for:</strong> A smooth and flat plot suggests that the assumption is reasonably satisfied for that predictor</td>
</tr>
</tbody>
</table>
Appendix
Overview of Maximum Likelihood Estimation of a Cox PH Model

The Cox PH model is a semi-parametric model. Its estimation considers what is called a partial likelihood.

- It is semi-parametric in that, while the role of the explanatory variables on survival is modeled explicitly, the baseline hazard is allowed to be arbitrary.

\[
\begin{align*}
h(t; X_1, \ldots, X_p) &= h_0(t) \exp[ \beta_1 X_1 + \ldots + \beta_p X_p ] \\
&= \text{Parametric.} \\
&= \text{Model free.}
\end{align*}
\]

- The partial likelihood approach to its estimation involves ideas similar to those we considered in obtaining the Kaplan-Meier estimates of survival. Specifically, we work with a series of conditional likelihoods each of which is a consideration of who is at risk just before each actual occasion of event.

Overview of the partial likelihood approach to estimation.

- A full detail is beyond the scope of this course.

- The setting is the following:

1. \( N \) = total number of subjects who enter the study at time zero.
2. \( n \) = number of actual events (deaths, whatever)
3. \( t_1 < t_2 < \cdots < t_n \) are the ordered times of the actual events.
- The instants “just before” actual events are $t_1^- < t_2^- < \ldots < t_n^-$

- The pattern of censoring is this:

  (1) Between $[0, t_1)$: # censorings is $C_1$
  (2) Between $[t_1, t_2)$: # censorings is $C_2$
    \hspace{0.5cm} etc \hspace{0.5cm} etc
  (i) Between $[t_{i-1}, t_i)$: # censorings is $C_i$
    \hspace{0.5cm} etc \hspace{0.5cm} etc
  (n) Between $[t_{n-1}, t_n)$: # censorings is $C_n$

- At each of these “just before” instants, we define a “risk set”. These are persons who have: (1) not yet experienced an event; and (2) have not yet been censored. They are still at risk. Thus, these persons had to have survived this long in order to be still at risk.

  (1) $R(t_i) =$ the risk set at time $t_i^-$
    \hspace{0.5cm} = \text{collection of individuals who have survival or censoring times} \geq t_i$
  (2) $\# R(t_i) =$ the number (count) who are still at risk (the risk set) at time $t_i^-$

- The full likelihood (if we were to use it) is the product of two types of “incremental” (actually conditional) likelihoods. This is again the idea of thinking in stages (just as we did in Kaplan Meier estimation methodology).

  (1) Censoring: Likelihood of $C_i$ censures & NO failures in $(t_{i-1}, t_i)$ given history up to time $t_{i-1}$.
  (2) Failure: Likelihood of failure at time $t_i$ given history up to “just before” time $t_i$.

<table>
<thead>
<tr>
<th>$c_1$ censored</th>
<th>$c_2$ censored</th>
<th>$c_n$ censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 N</td>
<td>$t_1$</td>
<td>$t_n$</td>
</tr>
<tr>
<td>Start</td>
<td>1&quot; event</td>
<td>$n^{th}$ event</td>
</tr>
<tr>
<td></td>
<td>$#R(t_1) = r_1$</td>
<td>$#R(t_n) = r_n$</td>
</tr>
</tbody>
</table>

- \hspace{0.5cm} etc \hspace{0.5cm} etc
- **Likelihood (full data)** is the product of several “incremental” conditional likelihoods as follows.

\[
L_{\text{FULL}} = \text{likelihood } [ C_1 \text{ censored in } (0, t_1) \text{ and no failures in } (0, t_1) ] \\
* \text{likelihood } [1^{\text{st}} \text{ failure at } t_1 \mid \text{history to time } t_1 ] \\
* \text{likelihood } [ C_2 \text{ censored in } (t_1, t_2) \text{ and no failures in } (t_1, t_2) \mid \text{history at } t_1 ] \\
* \text{likelihood } [2^{\text{nd}} \text{ failure at } t_2 \mid \text{history to time } t_2 ] \\
\text{Etc etc} \\
* \text{likelihood } [ C_n \text{ censored in } (t_{n-1}, t_n) \text{ and no failures in } (t_{n-1}, t_n) \mid \text{history at } t_{n-1} ] \\
* \text{likelihood } [ n^{\text{th}} \text{ failure at } t_n \mid \text{history to time } t_n ] \\
* \text{likelihood } [ \text{everyone else censored at } t_n \mid \text{history at } t_n ]
\]

- **Note that there are two types of terms in this product:**
  1. A likelihood of **censorings**
  2. A likelihood of events (**failures**)

- **The partial likelihood** argument reasons as follows:

  1. The conditional likelihoods for the censures contain NO information about the explanatory variables \(X_1, X_2, \ldots, X_p\).
  2. Since we are assuming that censoring is independent of \(X_1, X_2, \ldots, X_p\)
  3. No information is lost by dropping the conditional censoring likelihoods
  4. It is okay to maximize the product of the conditional failure likelihoods. This product is called the **partial likelihood of the data**.

- **Partial Likelihood** is also a product, but it considers only the actual event times, not the censorings.

\[
L_{\text{PARTIAL}} = \text{likelihood } [1^{\text{st}} \text{ failure at } t_1 \mid \text{history to time } t_1 ] \\
* \text{likelihood } [2^{\text{nd}} \text{ failure at } t_2 \mid \text{history to time } t_2 ] \\
\text{Etc etc} \\
* \text{likelihood } [ n^{\text{th}} \text{ failure at } t_n \mid \text{history to time } t_n ]
\]

Now we make use of the product notation (this is similar to the summation notation)

\[
L_{\text{PARTIAL}} = \prod_{i}^{n \text{ events}} \text{likelihood } [i^{\text{th}} \text{ fails at time } t_i \mid \text{history to } t_i, i^{\text{th}} \text{ survives to time } t_i ]
\]
Notice that these conditional likelihood expressions are exactly what we mean by a hazard function. This means that the above expression is actually the following:

\[
\frac{n \text{ events}}{\prod_i \text{likelihood} \left[ \{i^{th} \text{ fails at time } t_i \} \mid R_i \right]} \frac{\text{likelihood} \left[ \text{exactly 1 failure at time } t_i \mid R_i \right]}{\text{likelihood} \left[ \text{exactly 1 failure at time } t_i \mid R_i, \text{ survival up to time } t_i \right]}
\]

Terrific!! We have a model for the hazard function!! It is the Cox proportional hazard model of the hazard function. Thus, under the Cox PH model the above is:

\[
\frac{n \text{ events}}{\prod_i \left( \frac{h(t_i; X_{i1}, X_{i2}, \ldots, X_{ip})}{\sum \text{ all individuals } "i" \text{ in risk set } R_i h(t_i; X_{j1}, X_{j2}, \ldots, X_{jp})} \right) = \prod_i \left( \frac{h_0(t_i) \exp(X_{i1} \beta_1 + X_{i2} \beta_2 + \ldots + X_{ip} \beta_p)}{\sum \text{ all } "i" \text{ in } R_i h_0(t_i) \exp(X_{j1} \beta_1 + X_{j2} \beta_2 + \ldots + X_{jp} \beta_p)} \right) = \prod_i \left( \frac{\exp(X_{i1} \beta_1 + X_{i2} \beta_2 + \ldots + X_{ip} \beta_p)}{\sum \text{ all } "i" \text{ in } R_i \exp(X_{j1} \beta_1 + X_{j2} \beta_2 + \ldots + X_{jp} \beta_p)} \right)
\]

\[ Remarks \]

1. **This is the basis of estimation and it is an application of maximum likelihood estimation.** In brief, estimation involves some calculus (Newton Raphson iteration) which seeks to find choices of the betas that make as large as possible the magnitude of this partial likelihood.

2. This approach is appropriate **ONLY IF** the pattern of censoring contains NO information about the betas.